# **< DRAFT TITLE OF THESIS >**

**Ali Suhail**

**Supervised by: Felipe Campelo**

**Abstract: <** a paragraph or two that briefly outlines the motivation of your work, what you will do, and a "fairy tale" outcome of what you hope to find.

Student dropout remains a significant challenge in higher education, impacting both individual academic outcomes and institutional performance. Early identification of at-risk students is crucial for enabling timely interventions that improve retention. This project aims to develop a machine learning (ML) based dropout prediction system that can accurately classify students as likely to continue or withdraw, using demographic, academic, and behavioural data available at specific points during a module.

The approach involves constructing a full ML pipeline, beginning with exploratory data analysis and data cleaning, followed by model development and evaluation. Multiple algorithms including logistic regression, support vector machines, random forests, and neural networks will be trained, optimised, and compared using key performance metrics. The objective is to identify the most effective and interpretable models for this classification task.

The ideal outcome is a robust, interpretable predictor capable of flagging at-risk students early with high accuracy. Beyond prediction, the project seeks to highlight key factors contributing to dropout, empowering educators to take data-driven actions that enhance student retention and academic success.

**Ethics statement:** This project fits within the scope of the blanket ethics application, as reviewed by my supervisor Felipe Campelo. I have completed the ethics test on Blackboard. My score is 12/12.

**Project plan:**

**< t**wo pages of coherent text (i.e., of the form you intend to write for your thesis, not rough notes) that could be used as the opening pages of your Introduction/Overview chapter (Chapter 1 of your thesis).Describe what you plan to do, why it is interesting, and what you hope to discover.

Student dropout is a persistent challenge in higher education, affecting not only the academic success of individuals but also the financial stability and reputation of educational institutions. Identifying students at risk of discontinuing their studies early can enable timely interventions, improving retention rates and student outcomes. However, predicting dropout is complex due to the multifaceted nature of the factors involved, including demographic, academic, and behavioural elements. The motivation for this project is to develop a reliable and scalable student dropout prediction system that harnesses machine learning (ML) techniques to help educators and administrators understand and mitigate dropout risks more effectively.

---

Why is it interesting?

The rising dropout rates in higher education have become an increasing concern globally, with serious implications for students, educational institutions, and policymakers. Although greater access to university education has created a larger pool of graduates for the labour market, it has also resulted in a notable increase in the number of students leaving before completing their degrees [1]. According to the OECD (2019), dropout rates are increasing by an average of around 30 percent across many countries [2]. This highlights the need for effective strategies to identify and support students who are at risk of disengaging, while still maintaining academic standards despite growing enrolment figures.

In the United Kingdom, the problem is reflected in data from the Student Loans Company (SLC), which shows a 28 percent increase in university dropouts over a five-year period. The number of students who took out loans but did not complete their courses rose from 32,491 in 2018–19 to 41,630 in 2022–23 [3]. Mental health difficulties have been identified as a leading reason for early withdrawal [3]. These findings underline the importance of early intervention and predictive tools that can alert academic staff to students who may need extra support. Prior studies demonstrated the effectiveness of targeted academic interventions like personalised emails and proactive tutor engagement [4]. Their work showed an 11 percent reduction in dropout rates within classes that received such interventions [4]. Although the study acknowledged that other factors like course design may influence outcomes, it did not explore these in detail. It also noted that distance learning offers valuable opportunities for tracking and responding to student engagement.

Predicting student dropout is also important for managing academic resources and improving learning outcomes. Accurate predictions allow institutions to provide timely support, such as tutoring or customised learning pathways. They also enable better planning, such as adjusting teaching staff levels or identifying courses that may need revision. By implementing machine learning models that forecast dropout risk, universities can take data-driven actions to reduce attrition, improve retention, and enhance the overall quality of education.

---

This project involves constructing a comprehensive ML pipeline tailored for dropout prediction. Initially, exploratory data analysis (EDA) will be performed to uncover key patterns and relationships within the student data, including demographics, academic performance, and virtual learning engagement metrics. Rigorous data preprocessing and cleaning will be carried out to handle missing values, outliers, and categorical variable encoding. The dataset will then be split into training and testing sets to ensure unbiased evaluation. Multiple ML models, including classical algorithms like logistic regression, support vector machines, random forests as well as neural network models, will be developed and optimised through hyperparameter tuning. These models will be compared based on performance metrics such as accuracy, precision, recall, and F1-score, providing a clear ranking and insight into their relative strengths for this prediction task.

The ideal outcome for this project is the development of a reliable and interpretable ML-based dropout prediction system that can effectively identify students at risk of leaving their studies, based on the data available at a specific point during the module timeline. The task is formulated as a binary classification problem, where the model outputs 0 (student is likely to continue) or 1 (student is likely to drop out). In addition to achieving strong predictive performance, the project also seeks to uncover the most influential factors associated with student dropout, providing actionable insights for academic support teams. In a best-case scenario, this tool would not only deliver accurate predictions but also support educational institutions in making proactive, data-driven decisions to improve retention and promote student success and wellbeing.

MAX TWO PAGES >

**References:**

< A list of all the literature sources cited in your literature survey, **consistently**formatted in a commonly-used style (such as APA or IEEE), and with each item in the References being **complete**, i.e. as you would format it in your final submitted thesis, with author names, publication year, publication venue (conference or journal name), page numbers, DOI, etc.

References do not count towards your total page count

[1] “Factors influencing academic performance and dropout rates in higher education,” *Oxford Review of Education*, 2025, doi: https://doi.org/10.1080//03054985.2024.2316616.

‌

[2] “Executive summary,” *OECD*, 2025. https://www.oecd.org/en/publications/education-at-a-glance-2019\_f8d7880d-en/full-report/component-5.html#execsumm-d1e1370 (accessed Jun. 13, 2025).

‌

[3] J. Bryson, “University dropout rates reach new high, figures suggest,” *BBC News*, Sep. 28, 2023. https://www.bbc.co.uk/news/education-66940041 (accessed Jun. 13, 2025).

‌

‌

[4] “A systematic review on the deployment and effectiveness of data analytics in higher education to improve student outcomes,” *Assessment & Evaluation in Higher Education*, 2020, doi: https://doi.org/10.1080//02602938.2019.1696945.

‌

>

**Appendix: Project Timeline**

**<** one-page time-plan for your project, which you may choose to format as a week-by-week bullet-list, or possibly as a Gantt Chart

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TASK** | **PRIORITY** | **PROGRESS** | **START** | **END** |
|  |
| **Project Initiation** |  |  |  |  |  |
| Read Project Description & Literature | Must Have | 100% | 02/06/2025 | 06/06/2025 |  |
| Gantt Chart & Identify Risks (MoSCoW) | Must Have | 5% | 06/06/2025 | 08/06/2025 |  |
| Setup Trello Kanban & GitHub Repository | Must Have | 100% | 06/06/2025 | 09/06/2025 |  |
| Ethics Review & Test | Must Have | 100% | 06/06/2025 | 09/06/2025 |  |
| Project Plan | Must Have | 5% | 10/06/2025 | 22/06/2025 |  |
| **Data Exploration and Preprocessing** | | | | |  |
| Exploratory Data Analysis (EDA) | Must Have | 44% | 14/06/2025 | 17/06/2025 |  |
| Data Cleaning & Preprocessing | Must Have | 0% | 18/06/2025 | 21/06/2025 |  |
| Establish Machine Learning (ML) Pipeline | Must Have | 0% | 22/06/2025 | 24/06/2025 |  |
| Train/Test Split | Must Have | 0% | 25/06/2025 | 26/06/2025 |  |
| Data Scaling & Transformation | Should Have | 0% | 27/06/2025 | 29/06/2025 |  |
| **Model Development & Evaluation** | | | | |  |
| Run Baseline ML Models | Must Have | 0% | 30/06/2025 | 05/07/2025 |  |
| Model Tuning & Optimisation | Should Have | 0% | 06/07/2025 | 12/07/2025 |  |
| Evaluate Model Results | Must Have | 0% | 13/07/2025 | 17/07/2025 |  |
| Analyse & Interpret Results | Must Have | 0% | 18/07/2025 | 24/07/2025 |  |
| Real-time Dropout Prediction Dashboard | Could Have | 0% | 24/07/2025 | 28/07/2025 |  |
| **Analysis and Reporting** |  |  |  |  |  |
| Write Results Section | Must Have | 0% | 25/07/2025 | 31/07/2025 |  |
| Write Methodology Section | Must Have | 0% | 01/08/2025 | 07/08/2025 |  |
| Write Introduction & Literature Review | Must Have | 0% | 08/08/2025 | 14/08/2025 |  |
| Write Conclusion & Abstract | Must Have | 0% | 15/08/2025 | 21/08/2025 |  |
| Write Future Work (Optional) | Could Have | 0% | 17/08/2025 | 21/08/2025 |  |
| Final Edits & Proofreading | Must Have | 0% | 22/08/2025 | 27/08/2025 |  |

**Appendix: Risk Assessment**

< one-page risk assessment for your project, talking about the major risks you can foresee that might plausibly occur and interfere with your plans. For each risk, state clearly what it is, what its likelihood is, what its effects/impact would be on the project, and what your intended mitigation or risk-reduction involves.

|  |  |  |  |
| --- | --- | --- | --- |
| Risk | Likelihood | Impact | Mitigation |
| Data Quality Issues | Medium |  |  |
|  |  |  |  |
|  |  |  |  |

MAX ONE PAGE

>